IRSTI 16.21.33 https://doi.org/10.26577/EJPh.2024.v193.i1.ph2

**A.B. Amirbekova[1]* , G.M. Mamyrbek[1] , G. Talgatqyzy[1] , Urakova Yanch L.[2]**

[1]A. Baitursynov Institute of Linguistics, Kazakhstan, Almaty
[2]Akdeniz University, Turkey, Antalya
*e-mail: marghan01@mail.ru

# WAYS TO DEVELOP A SUBCORP OF THE WRITER'S TEXT: STRUCTURE AND FUNCTIONS OF META-MARKUP

Currently, based on the achievements of the world language corpus, the improvement of the national corpus of the Kazakh language, including subcorps, is becoming relevant. The article is devoted to the discussion of the principles and models of the development of the subcorpus of the writer's text. The subcorpus of the writer's text is an annotation base of literary texts in the genre of prose and drama. The subcorpus of the writer's text represents the base of visual means and tropes expressing the writer's artistic language. This resource allows the user to get detailed information about the writer, about his work, about a work of fiction (novel, novella, short story, short story, play), as well as read the work online. The national corpus of the Kazakh language was created on the basis of a search engine. The subcorp of the writer's text is also developed using three search engines (author, keyword, work). In this regard, the article analyzes the content of meta tags, which provides information about the author and the artwork.

Models of the introduction of linguistic research into the corpus base serve as the basis for the development of semantic markup of visual means such as metaphor, comparison, epithet, etc. in the corpus of the writer. The article analyzes the subcorpus structures within the framework of world computational linguistics and provides a sample of the meta-markup of an artistic work in the national corpus of the Kazakh language. The subcorpus of the writer makes it possible to search for information in literary texts – to select the text (texts) by name, by author, by gender of the author, by year of birth of the author, by year of creation of the text, all artistic prose texts, a separate genre of fiction, a separate type of literary text, texts in accordance with the place and time of the events described. The subcorpus includes works of fiction, that is, prose and dramatic texts. The study used the EXMARaLDA software package, the HIAT software method, as well as linguistic stylistics methods. Genre types of prose and drama were highlighted. The subjects of the works of art were determined. The classroom age of each work of art has been determined.

**Key words:** writer's subcorpus, annotation meta-markup, semantic meta-markup, model, text base, prose, drama.

А.Б. Әмірбекова[1]*, Г.М. Мамырбек[1], Г. Талғатқызы[1], Уракова Янч Л.[2]

[1]А. Байтұрсынұлы атындағы Тіл білімі институты, Қазақстан, Алматы қ.
[2]Акдениз университеті, Түркия, Анталия қ.
*e-mail: marghan01@mail.ru

## Жазушы мәтіні ішкорпусын әзірлеу жолдары: метабелгіленім құрылымы мен қызметтері

Қазіргі таңда әлемдік тіл корпусының жетістіктеріне сүйене отырып, қазақ тілінің ұлттық корпусын, соның ішіндегі ішкорпустарды жетілдіру өзекті болып отыр. Мақала ҚТҰК-ның жазушы мәтіні ішкорпусын әзірлеудің алғышарттарын талқылауға арналады. Жазушы мәтіні ішкорпусы – проза, драма жанрындағы көркем шығарма мәтіндерінің аннотацияланған smart қоры және жазушының көркем тілін танытатын көріктеуіш құралдардың базасы. Бұл ресурс пайдаланушыға жазушы туралы мәлімет, сонымен бірге роман, хикаят, повесть, әңгіме, мысал, новелла, пьеса, т.б. жанрдағы қазақ көркем сөз туындысы жайында толық ақпарат алуға, сол шығарманы онлайн оқуға мүмкіндік береді. Қазақ тілінің ұлттық корпусы іздеу жүйесіне негізделген. Жазушы мәтіні ішкорпусы да үш іздеу жүйесімен (автормен, кілт сөзбен және шығарма атауымен) әзірленеді. Осыған орай мақалада автор мен көркем туынды туралы мәлімет беретін белгіленімдердің мазмұны талданды.

Корпустық базаға тілтанымдық ілімдерді енгізудің модельдері жазушы ішкорпусындағы көркем шығармаға рәсімделетін метабелгіленім мен әрбір көріктеуіш сөздерге берілетін семантикалық белгіленім үлгілерін талап етеді. Мақалада қазақ тілінің ұлттық корпусы құрамындағы жазушы ішкорусының құрылымы, әлемдік компьютерлік лингвистика аясында талданды. Жазушы мәтіні ішкорпусы көркем мәтіндерден ақпарат іздеуге – мәтін тақырыбы, авторы, автордың жынысы, автордың туған жылы, шығарманың жазылған жылы, барлық көркем проза жанры, жеке жанрлық түрі, шығармадағы оқиғалардың орны мен уақыты туралы мәліметті алуға мүмкіндік береді.

Зерттеуде EXMARaLDA бағдарламалық жиынтығы, HIAT бағдарламалық әдісі, сондай-ақ лингвостилистикалық әдіс қолданылады. Проза және драмалық көркем шығармалардың жанрлық түрлері ажыратылады. Қазақ көркем туындыларының тақырыптары толықтырылды. Әрбір көркем шығармаға тән аудиториялық жас ерекшелігі анықталады.

**Түйін сөздер:** жазушы ішкорпусы, аннотациялық метабелгіленім, семантикалық белгіленім, модель, мәтін базасы, проза, драма.

А.Б. Амирбекова[1]*, Г.М. Мамырбек[1], Г. Талгатқызы[1], Уракова Янч Л.[2]

[1]Институт языкознания имени А. Байтурсынулы, Казахстан, г. Алматы
[2]Университет Акдениз, Турция, г. Анталья
*e-mail: marghan01@mail.ru

## Приемы разработки подкорпуса текста писателя: структура и функции метаразметки

В настоящее время, основываясь на достижениях мирового языкового корпуса, актуальным становится совершенствование национального корпуса казахского языка, в том числе подкорпусов. Статья посвящена обсуждению принципов и моделей развития подкорпуса текста писателя. Подкорпус текста писателя представляет собой аннотационную базу художественных текстов в жанре прозы, драмы. Подкорпус текста писателя представляет базу изобразительных средств и тропов, выражающих художественный язык писателя. Этот ресурс позволяет пользователю получить подробную информацию о писателе, о его творчестве, о художественном произведении (роман, повесть, рассказ, новелла, пьеса), а также прочитать произведение онлайн. Национальный корпус казахского языка создан на основе поисковой системы. Подкорпус текста писателя также разрабатывается с помощью трех поисковых систем (автор, ключевое слово, произведение). В связи с этим в статье анализируется содержание метатегов, на которой представлена информация об авторе и художественном произведении.

Модели внедрения лингвистических исследований в корпусную базу служат основой для разработки семантической разметок изобразительных средств, как метафора, сравнение, эпитет, и.т.д в корпусе писателя. В статье анализируется структуры подкорпуса в рамках мировой компьютерной лингвистики и дается образец метаразметки художественного произведения в национальном корпуса казахского языка. Подкорпус писателя дает возможность поиска информации в художественных текстах — отобрать текст (тексты) по названию, по автору, по полу автора, по году рождения автора, по году создания текста, все художественные прозаические тексты, отдельный жанр художественной прозы, отдельный тип художественного текста, тексты в соответствии с местом и временем описываемых событий. Подкорпус включает в себе художественных произведении, то есть прозаических и драматических текстов. В исследовании использовались программный пакет EXMARaLDA, программный метод HIAT, а также методы лингвостилистики. Были выделены жанровые типы прозы и драмы. Были определены тематика художественных произведений. Определен аудиторный возраст каждого художественного произведения.

**Ключевые слова:** подкорпус текста писателя, аннотационная метаразметка, семантическая метаразметка, модель, текстовая база, проза, драматургия.

## Introduction

In linguistics, a corpus (corpora) is a set of texts selected and edited according to certain rules that serve as the basis for language research. They are used for statistical analysis and verification of statistical hypotheses confirming linguistic rules in a given language. The corpus of texts is the subject of the study of corpus linguistics. The main properties of the text corpus can be expressed as follows:

– electronic text database – electronic text design;

– representative text base – comprehensive representation of the simulated object;

– fixed text base – transmission of the main difference between the corpus and the collection of texts.

– pragmatically oriented text base – subcorps created for a specific task (Kolpakova, 2012: 76).

The purpose of the study of the subcorp of the writer's text, which is part of the corpus, is to determine what features are present in the meta–markup of works of fiction, including works in the genre of prose and drama based on the texts of the corpus, its structure and functions. The interest in choosing this topic is to introduce readers to the works of writers and promote online reading of works.

The relevance of the development of the writer's subcorpus lies in the selection of artistic works into a thematic group through the digitalization of the literary text, the definition of thematic genres of Kazakh prose, and the determination of the level of works in demand by society.

The subcorp of the writer's language text is an annotated electronic collection of works of fiction in the genres of prose and drama. That is, through the text, the writer gets the opportunity to get full information about the Kazakh novel, novella, novella, novel, example, short story, work in the genre of the play and read it online. This interface aims to be user-friendly for most readers. Therefore, when compiling the subcorpus system, the principles of collecting the writer's works and sorting them in a certain sequence were determined. Let's see how the work on the development of the writer's subcorpus was conducted:

In order to improve the national corpus of the Kazakh language (https://qazcorpus.kz/) for the first time, the preparation of the database of "Texts of the writer's language" was started.

Since the corpus of texts is an algorithmic program that combines these texts with a certain logical system, the author's text was developed based on the following principles:

1. The corpus of texts is classified as dynamic and static. The inner body of the writer's text is continuously replenished. Because the artistic works of new young writers are being added. Therefore, the writer's text was developed in accordance with the principle of a dynamic corpus.
2. The corpus of texts is divided into two depending on the completeness and fragmentary nature of the text. Their difference will depend on the size. The inner body of the writer's text corresponded to a full-text, that is, a three-dimensional corpus format.
3. The corpus of texts is divided into the corpus of general texts and texts of individual authors. The inner body of the writer's text was modeled after the texts as a whole.
4. The corpus of texts is divided into research or illustrative (for example material) depending on the direction of the user. In this regard, the writer's text had to be compiled as a basis for obtaining illustrative material.
5. The corpus of texts differs by types of notation (metacognition, linguistic notation, prosodic, semantic, morphological notation). The inner body of the writer's text was developed in the format of a meta-designation.
6. The corpus of texts, depending on the consumer ability of the reader, is divided into open access, commercial, closed (with preferred consumers). The internal corpus of the writer's text is being developed on an open-access IT resource.
7. The corpus of texts is developed in trilingual, bilingual, monolingual, corresponding to the user's language. In this regard, the inner body of the writer's text is being developed in the Kazakh language.
8. The corpus of texts, depending on the direction of the search, differs in searching by lines with the author, keyword, title of the work, word, appendices, root, etc. The inner body of the writer's text is developed by three search engines (the author, the keyword, the title of the work).

The writer's text fund has been developed in accordance with the "corpus manager" format. A case manager is a software device designed to provide a text resource with a specialized search engine. Its function is to search for data in the corpus, output results to the user, and provide statistical information (Yatskevich, 2009:18).

To quickly search for a work of art selected by the reader, content was prepared for both systems so that the search through the author or the search by the title of the work of art would be performed in parallel.

A chronological, periodic approach to the collection of works of fiction is applied to the collection of linguistic material necessary for the corpus base of texts of the writer's language.

1. Kazakh literary text of the XIX century (words of edification by Abay);
2. Kazakh literary text of the twentieth century
3. The artistic text of the XXI century.

In this sequence, the language materials necessary for the corpus database of the "Texts of the writer's language" were sorted.

## Material and methods

In the formation of the base of the writer's texts, first of all, the inductive and deductive method was used. First of all, popular works of classical writers were collected. In the work on sorting them into thematic groups, the method of schematization was used. Also, the method of algorithmization was used in the selection of 100 Kazakh novels based on their ideological composition and chronological content. To be more specific:

First, the digitization method was used. Collected electronic version of literary texts. Each text was encoded in Latin letters and arranged in a certain order in a digitized system. The benefit of the digitization system is that it is preserved forever and is used at any time in the online network, on electronic resources. For example, the volume of more than 1,000 books in the database of the writer's texts, which contains 15 million words, is 700,000 Pages, is stored only in the size of one hard disk. It also doesn't get dusty and yellowed like a book. And most importantly, you can effectively use digitized text anywhere at any time. According to the research of E. V. Baranova: "In fact, digital humanities is a natural extension of the traditional field of humanities: These are digital methods that humanitarians use in their scientific activities, such as "3D modeling, computerized content analysis, databases, geoinformation systems" (Baranova, 2020: 46).

Secondly, the method of the EXMARaLDA software package was used. EXMARaLDA Corpus-Manager (Co-Ma) computer programs designed for special analysis processes. These types of programs are used for complex grammatical models. The work with the corps is carried out using special software tools – concordancers (a simpler type of programs) and corps managers, which provide various opportunities to obtain the necessary information from the corps (MacEnery, 2012:8). This method was used to insert the finished text into the corpus base, and it was used to easily insert an annotation meta-markup into each word application in the text. Thirdly, the corpus manager method was used. The corpus manager is a special search engine that uses software tools to search for data in the corpus, obtain statistical information and provide results to the user in a convenient form. The results of this procedure are presented as horizontal lines with a search word in the middle. This process is called KWIC (Key Word In Context) (Meyer, 2004: 12).

Fourth, the automation method was used. One of the functions of the automation method is the development of meta-markup. Linguistic markup is used to assign a code (tag) to a word, which denotes a set of grammatical features describing the word. Markup is conventionally divided into linguistic and externally linguistic markup. External linguistic ones include: – markup reflecting the peculiarities of text formatting (headings, paragraphs, indents, and so on); – markup related to information about the author and the text. The author may have his name, age, gender, years of life and others specified, and the text indicates the title, language, year and place of publication, and so on. Such information allows for a detailed search in the enclosures and provides tools to facilitate the identification of a particular document (Amieva, 2015: 252). New stylistic methods of Kazakh prose are shown in the article by Sh. Ismailova (Ismailova, 2023: 158).

## Literature review

Before developing the corpus of texts of the writer's language, the world experience was studied. The resource "Corpus of Narrative Prose of the XIX century", developed by the Russian Digital Research Laboratory (http://corpora.pushdom.ru/), is 500 works of fiction prose. The artistic text of the Russian classic writers of the XIX century is presented. This IT resource provides an opportunity to search and create concordances using a corpus of literary texts. The case is set to a continuous filling and improvement mode. But the meta-designation is not given. The resource "Children's Corpus", consisting of 1726 texts (A corpus of Russian prose for children and youth of the XX-XXI centuries) (http://pushkinskijdom.ru/), created for the purpose of online reading of a work of art. The meta description for the texts of the "Children's Corps" has not been compiled. As we have noticed, in world practice there was no meta-meaningful sample for the corpus of the writer's language text. Therefore, the "Meta-significant requirements and rules" developed based on the experience of the "National Corpus of the Kazakh language" were taken as a basis for the writer's text corpus. However, to determine the specifics of the text of the writer's language, research was conducted on the genre of prose.

In the Poetry Corpus (opened in 2006) (see 1) there are standard search types and non-standard search types – you can select texts not only according to the usual semantic and morphological markup, but also by genre, by meter, by stop, by clause, by stanza type, by rhyme type, and also for all the listed parameters. Poetic dramatic works have not yet been included in the corpus. It should be noted that the National Corpus is accompanied by a list and interpretation of the main poetic terms (Surovtseva, 2018: 421).

It is included in the subcorp of search texts – also in the corpus of biographical texts. "The corpus of biographical texts". The corpus was compiled to solve the problem of automatically searching for fragments containing biographical information in a natural language text (Glazkova, 2018: 224).

Another type of search text database – Russian General Internet Corpus (RGIC) is a searchable electronic online corpus of Russian texts from the Internet (webcorpora.ru). It was opened in 2013. The corpus includes text materials from the blogosphere, social networks, major news resources and literary magazines. The project has the status of an educational and scientific one, and many tasks of computational linguistics are solved by independent researchers and scientific groups based on the material obtained by the RGIC. While other corpus projects focus on fiction and edited texts, the General Internet Corpus provides Russian linguists with a timely opportunity to learn the language as it is, with all regional and slang features. The corpus makes it possible to carry out: Linguistic research of a wide range: dialectological research, the study of word distribution, the study of the language of social networks, the study of the influence of gender, age and other factors on language, the frequency of words, stable expressions and various constructions, stylistic features of texts from different segments of the Internet, and so on; Analysis of social networks; Machine learning based on the corpus, improvement of automatic markup algorithms (Piperski, 2013: 123).

The types of markup outside the corpus of texts and ways to develop it were also studied. In the world of corpus linguistics, the types of markup are divided into:

- morphological,
- semantic,
- syntactic,
- anaphoric,
- prosodic
- annotation.

Among them, semantic and annotation markup was used in the creation of the corpus of the writer's text. Therefore, we have given priority to the study of these named markings.

Semantic markup. Although there is no single semantic theory for semantics, semantic tags most often denote the semantic categories to which a given word or phrase belongs, and narrower subcategories specifying its meaning (Chesnokova, 2021: 6).

## Results and Discussion

The volume and genre types of language material collected on the corpus base of the "Texts of the writer's language" were diverse. In this regard, in accordance with the practice of developing a corpus of texts, works of art were selected by volume by sorting and storing depending on the volume of the text. Among the materials collected on the corpus base of the "Texts of the writer's language":

- novel – 83;
- novella – 396;
- story – 887;
- play – 21.

A total of 1,387 texts have been systematized.

The language materials collected in the corpus database of the "Texts of the writer's language" are collected in various formats (pdf, jpg, word). The texts are inserted into the corpus with the word version. Therefore, in the work on translating (converting) the collected texts into word format in various versions, the following activities were used https://pdf.io/r, website https://www.freepdfconvert.com/. All texts collected in the corpus database of the writer's texts were processed in the word version. In the process of editing in Word format, manual correction work was performed. Editing work was also performed in accordance with the Word size.

To systematize the work on collecting text in the corpus database of the "Texts of the writer's language", such a table was compiled.

**Table 1 –** Table of text collection in the Corpus database "Texts of the writer's language"

| № | The full name of the writer | Title of the work | The genre form of the work | | Lexicon |
|---|---|---|---|---|---|
| 1 | KALIKHAN ISKAK | It was late autumn | novella | 39 | 11 097 |
| 2 | | The meaning of the ambush | novella | 71 | 23 276 |
| 3 | | Fraction | story | 5 | 1 379 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | AKIM TARAZI | Blue House on a quiet street | comedy | 46 | 12 425 |
| 5 | | Good man | tragedy | 39 | 9 899 |
| 6 | | Lucky guy | drama | 44 | 12 180 |
| 7 | | Liner | tragicomedy | 42 | 11 431 |
| 8 | | Chandelier | farce-tragedy | 38 | 8 653 |

The corpus database of the “Texts of the Writer's Language” contains texts of works of fiction by 100 writers. In 2023, 1,387 texts were collected in the corpus database of “Texts of the writer's language” in the word version. The number of phrases of the writer's language texts is 8 million. Out of 8 million phrases, preparatory work was carried out for the linguistic designation of phrases (metaphor, comparison, epithet, phraseology, proverbs and sayings).

*Development of the corpus of texts of the writer's language.* After the accumulation of literary works by writers, the structural system of including authors in the composition of the writing team was determined. The interface for text input of the writer's language has been developed. One section of the national corpus of the Kazakh language – an example of the interface on the home page of the text internal corpus of the writer's text language was presented.

The artwork of each writer in the inner corpus of the writer's text language has been adjusted according to the online reading system. That is, the second part of the interface of the inner case was branched into genres of fiction:

1. The epic novel
2. The novel
3. The novella
4. The story
5. A short story
6. A fairy tale
7. The novel
8. The fable
9. The essay
10. Essays

One of the important issues identified in the inner body of the writer's language text is the style of the artwork. The style of artistic works is, of course, an artistic and literary style. Their internal style is defined: artistic and journalistic, artistic and historical. These internal styles have been incorporated into the meta-markup system. This is due to the fact that one of the main symbols defining the specifics of a writer is his inner style.

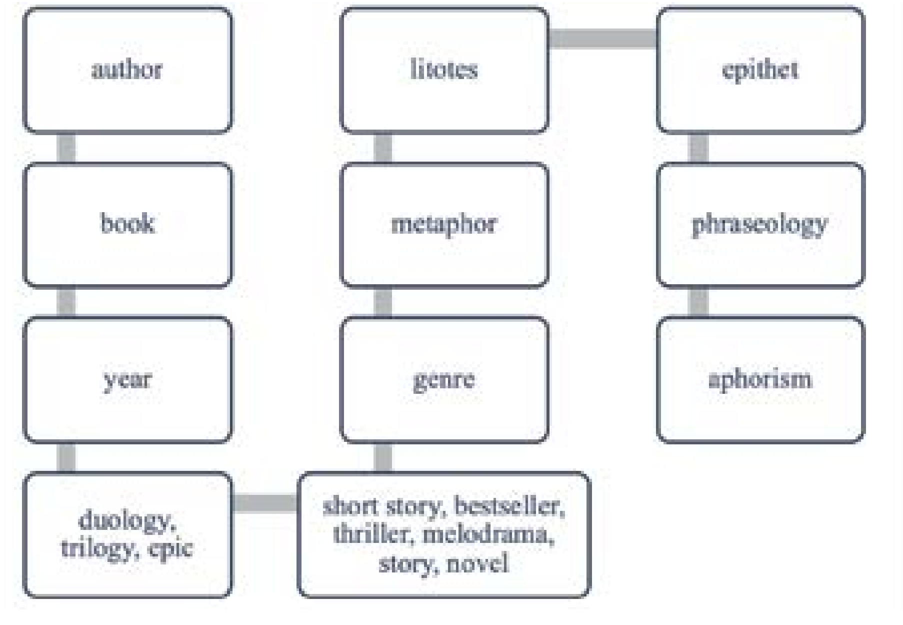

**Figure 1** – The second part of the interface of the internal corpus of the writer's language text

This interface provides information that defines who we call a writer, what is the difference between a writer and a poet.

A writer is the author of a work of fiction, that is, a text work. Usually the genres of writers are prose and drama. These include the authors of prose speech. Because prose speech (prose) is a large field of fiction. It differs from poetic poetry in that the rhythm of the word is free in it, corresponds to the norm of speech in literature, and any known criteria are not observed as in a poem.

In a work of fiction written in prose speech, a way of presenting the phenomena of life, various events is highlighted (Kazakh literature. Encyclopedic Reference, 2010: 89).

Prose is a literary genre, a work of fiction written in prose speech (short story, novella, novel). In prose, the phenomena of life and human characteristics are widely covered and comprehensively described (Kabdolov, 1992: 25). Since prose is an artistic text describing the life of society, the theme of the work was indicated in the inner body of the writer's text, taking into account the choice of the reader's interests. That is, the themes of the work:

– love;
– war;
– a way out of stress;
– the real life of the Soviet era;
– the famine period;
– the fate of Alash citizens;
– the relationship of mother-in-law and daughter-in-law;
– father-son relationship;
– the attitude of colleagues in the service;
– parenting;
– labor education;
– sports theme;
– the theme of art;
– the theme of death;
– childhood;
– youth;
– mother theme;
– social inequality;
– national liberation uprising;
– colonial theme;
– nostalgia for relatives;
– the theme of the native land, the fatherland;
– the topic of personal knowledge, political personalities;
– half a century of public life;
– the reconstruction period;
– the topic of virgin land development;
– the topic of production and industry;
– the communist period, etc.

These topics are grouped into a macro system. The meta-markups of the works illustrating new trends in Kazakh prose showed originality. For example: a grotesque story (T. Shapai's story "The Mistress of the House"), a social satire (K. Abulkhair's story "Paper City"). Thus, after the main features characteristic of the writer and the work of art were identified, an example of meta-markup was compiled.

**Table 2 –** A sample of the meta-markup of the internal corpus of the writer's language text

| **Meta markup of a writer's text subcorpus** | |
|---|---|
| **book** | |
| **author** | |
| **writer's autobiography** | |
| **summary of the book** | |
| **genre** | |
| **short story, bestseller, thriller, melodrama, story, novel** | |
| **chronotope** | |
| **style** | |
| **lexicon** | |
| **year, issue, publisher** | |

Before moving on to meta-tagging, the history of studying Kazakh prose was revealed. According to A. Baitursynov's research: "One of the talented words of "literary works" is a great story (novel); a long word (novella); and a story; a cheerful word; a fable; small stories" (Baitursynuly, 2005: 73).

If we talk about the fact that prose works originate in Turkic literature, then their further development covers the period of the XIII-XV centuries.

Along with the emergence of such concepts as oratory, correct speech, the genre of prose has become more active. This is due to the fact that one of the main tasks of prose is to reveal the affected topic with the help of an artistic image (adebiportal.kz).

The writers standing at the forefront of Kazakh prose include Ibrai Altynsarin, Abai Kunanbaevich, Mirzhakip Dulatov, Beimbet Mailin, Alikhan Bokeikhanov and many other major figures of Kazakh literature. The electronic version of fiction and drama by writers is fully assembled. A meta tag is made for each work. The works of these writers were borrowed specifically for the interface of the literary text of the twentieth century. Because this period was given in order to determine whether the Kazakh prose channel is the first developed, expanded one. Also included in the first interface were the stories of Ibray Altynsarin, who began prose works in an educational and didactic style. In the m-eta-markup of the first Kazakh novel, “The Unfortunate Zhamal” by Mirzhakyp Dulatov, it is noted that the work is also the first novel where the theme of female inequality in fiction was first described, and this novel is the beginning of the formation of Kazakh writers.

After determining the basic principles inherent in the Kazakh literary text, work was carried out on the meta-marking of the inner body of the text of the writer's language.

**Table 3 –** Development of meta-markup of the internal corpus of the writer's language text

| META-MARKING OF THE TEXT OF THE WRITER'S LANGUAGE | |
|---|---|
| Title of the work | The Red Arrow |
| The author of the work | Sherkhan Murtaza |
| Author's gender | Male |
| Biography, photo of the writer | Year of birth-death, place of birth, position, award for one famous work.<br>*For example:*<br>(1932 – 2018), was born in the Dzhambul region, Zhualy district, the village of Mynbulak. Public figure, Honored Worker of Culture of the Republic of Kazakhstan (1984), People's Writer of Kazakhstan (1992). In 1978, he received the State Prize of the Kazakh SSR for the novel “Black Coral”. |
| The main works of the writer | The novellas “The Found Sea” (1963), “The Son of an Unknown Soldier” (1969), “The Oath of Akhmetzhan” (1973), “The Unarmed Front”(1977), short stories “A Woman of 41 years” (1972), “Boarding School bread” (1974), novels “Black Coral” (1977), “The Red Arrow” from five books, “The Moon and Aisha” (1999), “There is always something missing in life” (2008), the plays “Letter to Stalin”, “Letter of Five”, “Who did not get into the reins” (“hero Bauyrzhan”), “mother Domalak” |
| The theme of the work | Human soul, love, war, theme of production, labor, national value, personification, **historical-revolutionary,** socio-historical event (optional, not all) |
| The epigraph of the work | *The Red Arrow of the Epoch*<br>*Monument to Turar Ryskulov*<br>Sh. Murtaza |
| A brief announcement of the work | The novel “Red Arrow” is one of the works written on a historical basis, an artistic memoir based on the life of Turar Ryskulov. |
| The first publication of the work | 1980 |
| The genre of the work | **prose,** drama |
| The genre form of the work | **novel, memoir** |
| The style of the work | Artistic and literary style |
| Chronotope of the work | The events of 1917-30. |
| Publisher, type of work | Dilogy, trilogy, tetralogy, two-volume, multi-volume.<br>Academic, scientific, public, etc. |
| A quote about the writer | In the new novel, Sherkhan grew up and became richer in terms of artistic speech. From words, pictures of nature, author's narrative colors corresponding to the character of the characters themselves, we learn such a culture of language as spaciousness, the sharpness of modern prose.<br>*Takhaoui Akhtanov* |

| | |
|---|---|
| The writer's Parables | A society without dreams is a society without hope<br>*Sh. Murtaza*<br>Democracy is not permissiveness,<br>Freedom is not haphazardness.<br>*Sh. Murtaza* |
| The artistic language of the work | (Metaphors, comparisons, phraseological units, proverbs and sayings)<br>To express submission to God (phraseology)<br>If the camel is shaken by the wind, look for a goat in the sky (proverb) |
| Number of phrases | 45004 |
| The source of the work | The red arrow. – Almaty, Publishing House “Writer”. 1980. – 256 pages. |
| The number of pages of the source of the work | 151 pages |
| Graphics of the source of the work | Cyrillic alphabet |
| Circulation | 2000 |
| The age of the audience | Teenagers, adults |
| The internal corpus | The writer's text |
| The source of the text | Literary Portal<br>https://adebiportal.kz/kz |
| The author who entered the work into the corpus, the time of entry into the corpus | Amirbekova A.<br>13/11/2023 |

The database of texts included in the internal corpus of the text of the writers of the national corpus of the Kazakh language consists of:

**Kazakh literary text of the XIX century**. 45 words of edification by Abay Kunanbayev.

**Kazakh literary text of the XX century.** Mukhtar Auezov's epic novel “The Path of Abai”, the unfinished novel “Prosperity”, the novella “The Story of Karash-Karash”, the story “The Harsh Century”, the novel by Sabit Mukanov “My Schools”, the novella “Baluan Sholak”, the novel “Syrdarya”, the play “Shokan Ualikhanov” and the novels “Bright Stars”, “Leaked star”, “Sulushash”, “Syrdarya”, “Listen, native wealth”, “Pure love”, “Years of growing up”, “School of Life”, “Botagoz”. The novella of Saken Seifullin “Diggers”, the novel “The Thorny Path”. Mirzhakyp Dulatov's novel “The Unfortunate Jamal”. Magzhan Zhumabayev's novel “The Sin of Sholpan”. Shakarim Kudaiberdiev's novel “Adil-Maria”. Sultanmakhmut Toraighyrov's novel “Beautiful Kamar”. Stories by Ibray Altynsarin “Abylai”, “There are many intelligent animals among them, but there is no complete mind like a human”, “The tale of the smart rich”, “Stupid friend”, “Golden Seeds”. Collection of short stories and fables by Spandiyar Kobeyev “Two plows”, novel “Bride Price”, collection of short stories and fables “Bird's Nest”, novel “A dream come true”. The novella of Zhusupbek Aimautov “Kunikey Wine”, the novel “Akbilek”, the novella “Kartkozha”, the story “Ghost”, the story “The Tragedy near Zhanabai”, the story “Tumarbai and his wife”, the story “The Saint predicting the blizzard”, the story “The Gardener and the boy”, the story “The Black Witch”, the story “The report of the Jew”, the story “On the road”, the story “The Singer”, the story “A giant dream”, the story “there were times, a pike in a pine head”, the story “Radio broadcasts”, the story “What is it if it doesn't change?”, the story “Leaves”, the story “Suitable pictures”, the story “Entered a new life”, the story “Nauryz in the steppe”, the story “When the train passed through the Urals”, the story “Blue Bull”, the story “What made Tinikey think”, the story “Nagima”, the story “Shepherd Tastamakh”, the story :How to eradicate illiteracy”, the story “Sop-sop”, the story “Stubborn Suley”, the story “Red Yurt”, the story “The Red Lady”, the story “Sokolnik”, the story “Oh, it's a matchmaker!”, the story “The Dragon”, the story “Nomadic Kuzhebai”, the story “Excited guy”. Stories by Beimbet Mailin “Black Bucket”, “Ulbosyn”, “The head of the dispute is the blue cow of Dayrabai”, “Moonish of Arystanbai”, “In the collective farm barn”, “House of the Red Army”, “The Story of Amirzhan”, the novellas “Fifteen houses”, “Good fellow”, “During threshing”, the novels “Azamat Azamatych”, “The Sign of Shuga”, “Raushan the Communist”, the novellas “Monstrous militants”, “Clippings”, “Front”, “When bypassing giants”, “Myrkymbai”. The novellas of Gabiden Mustafin “Fallen Rock”, “Revenge”, “Problems from the language”, “Proof”, “Prisone”, “The Man who did not laugh”, “Caravan”, ”Millionaire”, novels “Karaganda”, “After the Storm”, “Lookouts”. Novels by Gabit Musrepov

"Ulpan", "Kazakh soldier", "The Awakened Land", short stories "Autobiographical story", "In a stormy night", "Piglet", "Kusen". The novels of Zein Shashkin "Temirkazyk", "Breath of Life", "Akbota", the novels "Doctor Darkhanov", "Vera". Khamza Yesenzhanov's trilogy "White Zhaiyk", the novel "The Zhunusov Brothers", the novel "Many years later", the stories "The Old Man Kazakh", "When fishing", "Spouse". Taken Alimkulov's novel "The White Horse", the story "According to Seitek", the story "The Crimson River", the stories "In his native village", the story and collection of short stories "Kertolgau", the novel "The Fate of horses". Collection of short stories "Chronicle Sahara". Collections of short stories by Tolen Abdikov "The Unspoken Truth", the play "The Dead Bee", "There were three of us". Takhaoui Akhtanov's novel "Fierce Days", the story "Steppe Mystery", the novel "Buran", the dramas "Longing for love", "Unexpected meeting", the story "The Lion's Share", "Henpecked Husband".

The stories of Ilyas Yesenberlin "On the riverbank", "Yesil is boiling", "A song about a man". Novels "Fight", "Dangerous passage", "Lovers", "Rage", "Diamond Sword", "Golden Bird", "Commotion", "Protect with your shadow", trilogy "Nomads", trilogy "Golden Horde", novels "Holiday of Love", "Distant Islands", "The Joy of the Swan Bird". Novels by Abish Kekilbayuly "Noble", "The End of the legend", "Urker", "Twilight", novellas "Cloud", "Kui", "Jump horse", "Abyss", "Competition", "The Story of the Queen of Rivers", "Bird's Wing", "Jid bush", "Far the house". Novellas, short stories by Oralkhan Bokeya "Where are you, my foal", the story "Icy Mountain", a collection of novels and short stories "Urker departs", the novel "Do not extinguish your hearth", the story "Winter is long on our side", the story "I can't sleep", a collection of short stories. Mukhtar Magauin's short story "In the Evening", the novel "Rhythm of Kobyz", the novel "Blue Mirage", "Children of one grandfather", "Blue Mirage" (novels and short stories), historical novel-dilogy "Confusion", the novel "Shakan Sheri", autobiographical novel "I", collection of short stories "Massacre of ants".

**The literary text of the XXI century.** Ayagul Mantai's short stories "Tompak", "When the Birds Return", "Curse", "Blind Consciousness", "Writer", "Grave Flower". Nurgali Oraz's stories "The Path", "The children of Bakyrsh", "Kazygurt adventures", "Those who came from the steppe to the city", "Beauty at altitude", "Night solitude", "Holy Lake". Askar Suleimenov's novellas "After Noon", "Lost", "Besatar". Dramas "Horse and Rider", "Kulager", "Revenge", "Meeting with the writer", "The Seventh Chamber", novels "Face to face with reality", "Revenge".

## Conclusion

Summing up, when creating a corpus of texts by Kazakh writers of the XX -XXI centuries, the task is to include in the database of texts the maximum number of authors who wrote in the studied era, which will make linguistic research conducted on it more objective, since thereby not only the leading writers of the studied era are involved in the orbit of research, but also many secondary authors in the work in which all aspects of social and cultural life are represented to the fullest extent possible. The idea of systematic literature, in particular, the idea of a "literary and artistic system" that includes all literary and artistic products of a particular historical epoch.

The peculiarity of the created corpus is the genre of the story:

– A genre of rapid response that responds to the demands and challenges of the era and even sometimes anticipates them;

– Belonging to small forms of prose allows you to involve a large number of prose writers and their works into the orbit of research;

– Acute reaction to modernity – characteristic features of reality, features of the language of contemporaries;

– The richness and huge variety of individual style systems.

Tasks of the first stage of the corpus creation:

– formation of a representative list of authors and their works;

– cataloging of stories written and published during the period under study;

– search for electronic versions of texts;

– digitization of stories for which there are no electronic versions;

– proofreading of electronic versions of texts;

– meta tagging of short story texts;

– segmentation of texts into structural parts (sections, paragraphs, sentences);

– isolation of the author's text (narrative), characters' speech and author's remarks;

– linguistic annotation of texts.

Thus, the internal corpus of the text of the Kazakh literary language is assigned to each work. Therefore, the first designation in the table is the title of the work. The next sign is the author of the work.

Further designations are a complete introduction to the author of the abstract. The author's gender is given in order to determine the gender census. The main requirement was the introduction of a gender sign according to the rules of meta-marking of the national corpus of the Kazakh language. The procedure for transmitting information about the writer's life is defined. First, the years of birth, death, and locality. Secondly, the position, and thirdly, the awards for the best work. This information is recognized as the main features defining the image of the writer. Also, the main works of the writer were put as one sign. Because, interest is expected, if the reader wants to get acquainted with other works of the writer, he should be easily found. The following signs introduce the work of art. First, the theme of the work was set. If you like the theme of the work, then interest will grow through the following signs. The epigraph and brief abstract of the work required a brief and clear description (4-5 sentences) so that it would be understandable to the reader. This is followed by typographical information. It includes the first publication, circulation, edition, and source of the work. The following signs are artistic and stylistic signs of the work. These include genre, genre form, style of the work, artistic language and parables of the writer, and the number of words. The next feature was based on social interests. The indication of the age of the audience, depending on the age of the reader, contributes to the definition of children's literature, statistics of adult literature.

The corpus and database of writers' texts being developed will allow conducting research on the language and style of works of art, depending on a number of parameters – the social origin of the authors, their education, type of activity, the age of the writer at the time of writing a particular work.

**Support**

*This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant BR21882249 "Improving the tool of intercultural communication – the national corpus of the kazakh language (nckl) – and expanding its subcorpora"). We express our gratitude to the A. Baitursynov Institute of Linguistics, which supports the implementation of the project.*

**References**

Колпакова Г.В. Методы анализа корпусной лингвистики // Филологические науки. Вопросы теории и практики. – 2012. – №4(15). – С.75-77.

Қазақ тілінің ұлттық корпусы [Электронды ресурс]. – URL: https://qazcorpus.kz (Пайдаланылған күні: 20.12.2023)

Яскевич А.А. Корпусная лингвистика // Энциклопедия для школьников и студентов: в 12 т. Т.1: Информационное общество. XXI век. – Минск: Белорусская энциклопедия, 2009. – С. 167-169.

Баранова Е.В. Цифровая гуманитаристика: как историку и филологу понять программиста // РИА Новости, 9 июля, 2020 г. [Электронный ресурс]. – URL: https://na.ria.ru/20200709/1574061044.html (Дата обращения: 20.12.2023)

MacEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – NY: Cambridge University Press, 2012. – 254 p.

Meyer Ch.P. English Corpus Linguistics. An introduction. – NY: Cambridge University Press, 2004. – 198 p.

Амиева А.М., Филимонов В.В., Сергеев А.П., Тарасов Д. Инструменты корпусной лингвистики. – Екатеринбург: Уральский федеральный университет, 2022. – 233 с.

Лаборатория компьютерной лексикографии [Электронный ресурс]. – URL: https://philology-urgi.urfu.ru/ru/nauka/nauchnye-shkoly/uralskaja-semanticheskaja-shkola/laboratorija-kompjuternoi-leksikografii/ (Дата обращения: 20.12.2023)

Исмаилова Ш., Баянбаева Ж. Мифопоэтический анализ казахской повести второй половины ХХ века // Вестник КазНУ. Серия филологическая. – 2023. – №1 (189). – С. 157-163. https://doi.org/10.26577/EJPh.2023.v189.i1.ph1

Русская словесность. Корпуса текстов [Электронный ресурс]. – URL: http://corpora.pushdom.ru/narrative-prose.html (дата обращения: 20.12.2023)

Аспекты трансформации художественного текста [Электронный ресурс]. – URL: http://pushkinskijdom.ru/2023/05/16/aspekty-transformatsii-hudozhestvennogo-teksta/ (дата обращения: 20.12.2023)

Суровцева Е.В. Общеязыковые корпусы русского языка: подкорпусы художественных текстов // Молодой ученый. – 2018. – №49(235). – С. 420-422.

Глазкова А.В. Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке // Труды ИСП РАН. – 2018.Т.30. – Вып. 6. – С. 221-236. DOI: 10.15514/ISPRAS-2018-30(6)-12

Пиперский А.С. Общий интернет-корпус русского языка и понятие репрезентативности в корпусной лингвистике // Современные проблемы науки и образования. – 2013. – № 5. – Р. 120-140.

Чеснокова М.В. Виды корпусов текстов. – Тверь: Тверской государственный университет, 2021. – 12 с.

Қазақ әдебиеті: Энциклопедиялық анықтамалық. — Алматы: «Аруна Ltd.» ЖШС, 2010 – 254 б.

Қабдолов З. Сөз өнері. – Алматы: «Sanat», 2007. – 360 б.

Байтұрсынұлы А. *Әдебиет танытқыш. – Алматы*: Алаш, 2005. – 250 б.

Қазақ әдебиетіндегі проза жанрының даму сипаты [Электронды ресурс]. – URL: https://adebiportal.kz/kz/news/view/qazaq-adebietindegi-proza-zanrynyn-damu-sipaty__23618 (Пайдаланылған күні: 20.12.2023)

## References

Kolpakova, G. (2012). Metody analiza korpusnoi lingvistiki [Methods of analysis of corpus linguistics]. Filologicheskie nauki. Voprosy teori i praktiki [Philological sciences. Questions of theory and practice]. Iss. 4(15), P. 75-77. (in Russian)

Qazaq tilinin ulttyq corpusy [National corpus of the Kazakh language] [Electronic resource]. – URL: https://qazcorpus.kz (Date of use: 20.12.2023) (in Kazakh)

Yaskevich, A. (2009). Korpusnaia lingvistika [Corpus linguistics]. Ensiklopedia dlya shkolnikov i studentov [Encyclopedia for schoolchildren and students]. Minsk. Belarusian Encyclopedia. (in Belarus)

Baranova, E.V. (2020). Sifrovaia gumanitaristika: kak istoriku i filologu ponyat programista [Digital Humanities: how a historian and a philologist can understand a programmer]. RIA Novosti [RIA News]. [Electronic resource]. URL: https://na.ria.ru/20200709/1574061044.html (Date of use: 21.12.2023) (in Russian)

MacEnery, T., Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. NY. Cambridge University Press.

Meyer Ch. P. (2004). English Corpus Linguistics. An introduction. NY. Cambridge University Press.

Amieva, A.M., Filimonov, V.V., Sergeev, A.P., Tarasov, D.A. (2022). Sredstva keis-lingvistiki [Tools of case linguistics]. Ekaterinburg. Ural Federal University. (in Russian)

Laboratoria kompiuternoi lexsikografi [Laboratory of computer lexicography]. [Electronic resource]. URL: https://philology-urgi.urfu.ru/ru/nauka/nauchnye-shkoly/uralskaja-semanticheskaja-shkola/laboratorija-kompjuternoi-leksikografii/ (Date of the use: 20.12.2023) (in Russian)

Ismailova, Sh., Bayanbayeva, Zh. (2023). Mifopoeticheski analiz kazahskoi povesti vtoroi poloviny XX veka [Mythopoetical analysis of the kazakh story of the second half of the XX century]. Vestnik KazNU. Seria filologicheskaia [Bulletin of KazNU. The series is philological]. Iss.1(189), P. 157-163. Almaty. (in Kazakh)

Russkia slovesnost. Corpus texstov [Russian literature. Corpus of texts] [Electronic resource]. URL: http://corpora.pushdom.ru/narrative-prose.html (Date of use: 20.12.2023) (in Russian)

Aspekty transformatsi hudozhestvennogo texsta [Aspects of literary text transformation] [Electronic resource]. URL: http://pushkinskijdom.ru/2023/05/16/aspekty-transformatsii-hudozhestvennogo-teksta/ (Date of use: 21.12.2023) (in Russian)

Surovtseva, E.V. (2018). Obşeiazykovye korpusy ruskogo iazyka: podkorpusy hudojestvennyh tekstov [General linguistic corpus of the Russian language: under the corpus of literary texts]. Molodoj uchenyi [Young scientist], Iss. 49 (235), P. 420-422. (in Russian)

Glazkova, A.V. (2018). Avtomaticheski poisk fragmentov, soderjachih biograficheskuiu informasiu, v tekste na estestvenom iazyke [Automatic search for fragments containing biographical information in a natural language text]. Trudy ISP RAN [Proceedings of the ISP RAS], Iss. 6, P. 221-236. (in Russian)

Piperski, A.C. (2013). Obşi internet-korpus ruskogo iazyka i ponätie reprezentativnosti v korpusnoi lingvistike [The general internet corpus of russian and the notion of representativeness in corpus linguistics]. Sovremennye problemy nauki i obrazovania [Modern problems of science and education]. Iss. 4, P. 120-140. (in Russian)

Chesnokova, M.V. (2021). Vidy korpusov tekstov [Types of text corpora]. Tver. Tver State University. (in Russian)

Qazaq adebieti [Kazakh literature]. (2010) Ensiklopedialyq anyqtamalyq [An encyclopedic reference book]. Almaty. "Aruna Ltd." LLC. (in Kazakh)

Kabdolov, Z. (2007). Soz oneri [The art of the word]. Almaty. Sanat. (in Kazakh)

Baitursynuly, A. (2005). Adebiet tanytqysh [The Educator of literature]. Almaty. Alash. (in Kazakh)

Qazaq adebietındegı proza janrynyn damu sipaty [The nature of the development of the prose genre in Kazakh literature] [Electronic resource]. URL: https://adebiportal.kz/kz/news/view/qazaq-adebietindegi-proza-zanrynyn-damu-sipaty__23618 (Date of use: 20.12.2023) (in Kazakh)

**Information about authors:**

*1. Aigul Amirbekova (corresponding author) – Candidate of Philological Sciences, A. Baitursynov Institute of Linguistics (Almaty, Kazakhstan, email: marghan01@mail.ru);*

*2. Gulfar Mamyrbek– Candidate of Philological Sciences, A. Baitursynov Institute of Linguistics (Almaty, Kazakhstan, email: gulfar76@mail.ru);*

*3. Gulnara Talgatqyzy – Doctoral student, A. Baitursynov Institute of Linguistics (Almaty, Kazakhstan, email: gulalyzhan@mail.ru);*

*4. Lazzat Urakova Yanch – Docent, Akdeniz University (Antalya, Turkey, email: urakovayanc@gmail.com);*

Date of receipt of the article after correction: January 12, 2024.
Accepted: March 2, 2024.